THUNDR: Transformer-based 3D HUmaN Reconstruction with Markers

Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan
{mihaiz, andreiz, egbazavan}@google.com
William T. Freeman, Rahul Sukthankar, Cristian Sminchisescu
{wfreeman, sukthankar, sminchisescu}@google.com

Figure 1: Automatic 3d pose and shape reconstruction results with THUNDR. *(Left)* Input image. *(Middle)* Reconstructed 3d meshes projected on the camera plane and overlayed on the image. *(Right)* Different viewpoint showing our intermediate predicted marker representation (in green) and the reconstructed surface geometry. THUNDR provides automatic 3D scene placement of the reconstructed humans under a perspective camera model.

Abstract

We present THUNDR, a transformer-based deep neural network methodology to reconstruct the 3d pose and shape of people, given monocular RGB images. Key to our methodology is an intermediate 3d marker representation, where we aim to combine the predictive power of model-free-output architectures and the regularizing, anthropometrically-preserving properties of a statistical human surface model like GHUM—a recently introduced, expressive full body statistical 3d human model, trained end-to-end. Our novel transformer-based prediction pipeline can focus on image regions relevant to the task, supports self-supervised regimes, and ensures that solutions are consistent with human anthropometry. We show state-of-the-art results on Human3.6M and 3DPW, for both the fully-supervised and the self-supervised models, for the task of inferring 3d human shape, joint positions, and global translation. Moreover, we observe very solid 3d reconstruction performance for difficult human poses collected in the wild.

1. Introduction

The significant recent progress in 3d human sensing is supported by the development of statistical human surface models and the emergence of different forms of supervised and self-supervised visual inference methods. The use of statistical human pose and shape models offers advantages in the use of an anatomical and semantically meaningful human body representation, during both learning and inference. Human anthropometry could be used to regularize a learning and inference process, which, in the absence of such constraints, and given the ambiguity of 3d lifting from monocular images, could easily run haywire. This is especially true for unfamiliar and complex poses not previously seen in a 'training set'—as they never all are. Semantic models offer not only correspondences with image detector responses (specific body keypoints or semantic segmentation maps) which can give essential alignment signals for 3d self-supervision, but can also help rule out 3d solutions that may otherwise entirely break the symmetry of the body, the relative proportions of limbs, the consistency of the surface in terms of non self-intersection, or the anatomical joint angle limits.

The choice of evaluation metrics has an important role, too. For now, by far the most used representation—perceived as 'model-independent'—are the 'body joints', a popular concept, neither by virtue of its anatomical clarity (as that point idealization could be bio-mechanically argued against), nor—for computer vision, and more practically—given its lack of ground-truth observability. In practice, human 'body joints' are obtained either by fitting proprietary articulated

3d body models to marker data (internal models of the Mocap system, where the assumptions and error models are not always available) or by human annotators eye-balling joint positions in images, followed by multi-view triangulation to obtain pseudo-ground truth. While the latter have proven extremely useful in bootstraping initial 3d predictors, the joint-click positioning cannot be considered an accurate anatomical reality, in any single image, and even less so, consistently, over a large corpora, especially as for many non-frontal-parallel poses 'joint locations' are difficult to correctly identify, visually. While some form of 3d body joint prediction error seems unavoidable under the current ground-truth and state of the art metrics, a safeguard could be to operate primarily with visually grounded structures and obtain joint estimates using statistical body models, based on their surface estimates, as just a final step.

In this paper, we rely on the visual reality of 3d body surface markers (in some conceptualization, a 'model-free' representation) and that of a 3d statistical body (a 'model-based' concept) as pillars in designing a hybrid 3d visual learning and reconstruction pipeline. Markers have the additional advantages of being useful for registration between different parametric models, can be conveniently relied-upon for fitting, and can be used as a reduced representation of body shape and pose, as we will here show. Our model combines multiple novel transformer refinement stages for efficiency and localization of key predictive features, and relies on combining 'model-free' and 'model-based' losses for both accuracy and for results consistent with human anthropometry. Quantitative results in major benchmarks indicate state of the art performance. Extensive qualitative testing in the wild supports the overall feasibility, and the quality of 3d reconstructions produced by THUNDR, under both supervised and self-supervised regimes.

Related Work: There is considerable prior work in 3d human sensing which we only briefly mention here without aiming at a full literature review [30, 4, 23, 35, 28, 39, 29, 15, 16, 24, 6, 5, 42, 38]. Methods sometimes referred as 'model-based' [39, 16, 10, 27, 14, 3, 36, 41, 2, 40, 9] rely on statistical human body models like SMPL or GHUM, whereas others sometimes referred to as 'model-free' [32, 31, 13, 40, 20] rely on predicting a set of markers or mesh positions, without forms of statistical surface or kinematic regularization based on human anthropometry. While the second class of techniques tend to perform better in benchmarks (which are mostly emphasizing the prediction of 3d joint locations and occasionally joint angles), the former tend to be more semantically and anatomically intuitive, easier to deploy in the context of self-supervised learning, and more robust in environments, or for poses, not encountered during training. In this work we aim to leverage the advantages of both methods: predicting visually observable sets of markers, and yet regularize estimates using statistical kinematic pose and shape models. Moreover, additional innovations in the use of multiple layers of refining visual transformers, produce significant computational efficiency and accuracy gains in benchmarks, for self-supervised learning, and in the wild.

2. Methodology

In this section we review our methodology including the 3d statistical body models, the marker based-modeling, as well as the proposed THUNDR learning and inference architecture.

2.1. Statistical 3D Human Body Models

We use a recently introduced statistical 3d human body model called GHUM [35], to represent the pose and the shape of the human body. The model has been trained end-to-end, in a deep learning framework, using a large corpus of human shapes and motions. The model has generative body shape and facial expressions $\beta = (\beta_b, \beta_f)$ represented using deep variational auto-encoders and generative pose $\theta = (\theta_b, \theta_{lh}, \theta_{rh})$ for the body, left and right hands respectively represented as normalizing flows [37]. The pelvis translation and rotation are controlled separately, and represented by a 6d rotation representation [43] $\mathbf{r} \in \mathbb{R}^{6\times 1}$ and a translation vector $\mathbf{t} \in \mathbb{R}^{3\times 1}$ w.r.t the origin $(0, 0, 0)$. The mesh consists of of $N_v = 10,168$ vertices and $N_t = 20,332$ triangles. To pose the mesh, we apply the GHUM network $\mathbf{V}(\theta_b, \beta_b, \mathbf{r}, \mathbf{t}) \in \mathbb{R}^{N_v\times 3}$ to obtain the posed vertices. We omit the facial expressions and left and right hand poses, as we here focus on main body pose and shape. We also drop the b subscript for convenience.

Camera Model We adopt a pinhole camera model, with approximated intrinsics $\mathbf{C} = [f_x, f_y, c_x, c_y]^\top$ [38] and associated perspective projection operation $\mathbf{x}_{2d} = \Pi(\mathbf{x}_{3d}, \mathbf{C})$, where $\mathbf{x}_{3d} \in \mathbb{R}^{3\times 1}$. Because we work with cropped images, we also adapt our intrinsics, such that projecting the same 3d points – either in the cropped image or the original, full image – would give the same alignment. The transformation of image intrinsics $\mathbf{C}$ into corresponding crop intrinsics $\mathbf{C}_c$ is given by

$$[\mathbf{C}_c^\top 1]^\top = \mathbf{K}[\mathbf{C}^\top 1]^\top, \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{5\times 5}$ is the scale and translation matrix, adapting the image intrinsics $\mathbf{C}$.

By using a perspective camera model, we ensure that reconstructions are obtained in camera space. Hence, we have meaningful translation and relative positioning of one subject in the scene (or relative positioning of multiple subjects) when reconstructing from monocular images. A perspective model is a much more accurate and general representation of the imaging transformation compared to an orthographic one. It is in our view desirable in all cases, in order to go beyond showing just model projections or reconstructions in a human-centred coordinate system.

2.2. Marker-based Modelling

Current model-based architectures directly predict specific shape or pose parameters from a raw image. Inspired by model-free methods where weaker constraints are applied on outputs, we adopt an intermediate representation given by 3d surface markers. These can capture human shape and pose and we can predict them directly in a 3d camera space.

However, training purely model-free methods based on surface markers (as opposed to joint locations), faces additional challenges for both supervised and unsupervised learning. First as such markers are different from joint positions, very few datasets have labels for them. Markers may be available for motion capture datasets in both 2d and 3d, but training a reliable detector is not necessarily easy especially if one seeks generalization outside the lab, where most marker-based systems operate. For self-supervised learning, where additional forms of semantic (body part segmenation) analysis are often necessary, the lack of a statistical body model would render such potentially useful signals unavailable. Finally—and especially when learning with small supervised training sets or for exploratory self-supervised learning—, the lack of regularization given by a body model could lead to 3d predictions with inconsistent anthropometry, further derailing a convergent learning process.

Our approach is to use a 3d surface marker set as an intermediate representation proxy, controlled by both surface (mesh) properties and the parameters of a statistical 3d human pose and shape model (GHUM). For practical considerations, and without loss of generality, we adopt the Human3.6M marker set that consists of $N_m = 53$ units, see fig. 2 for details. We next describe two network heads, which given any 3d markers $\mathbf{M} \in \mathbb{R}^{N_m \times 3}$ achieve the following: *(i)* reconstruct the GHUM mesh through a simple architecture $\mathbf{V}_d(\mathbf{M}) \in \mathbb{R}^{N_v \times 3}$, and *(ii)* recover the corresponding GHUM parameters $(\theta, \beta, \mathbf{r}, \mathbf{t})$ from $\mathbf{M}$, so we can also recover an anthropometric mesh equivalent to $\mathbf{V}_d$, $\mathbf{V}_p(\theta, \beta, \mathbf{r}, \mathbf{t})$.

Training the Marker-based Poser (MP) The markers are essentially free 3 dof variables, but they follow the given surface placement description, in our case, the VICON protocol. To train a network that maps markers to vertices, we need a dataset of corresponding markers and vertices.

We take a synthetic sampling approach based on our GHUM model. Given generative codes for pose and shape $\theta, \beta \in \mathcal{N}(\mathbf{0}; I)$, $\mathbf{r}$ drawn from the Haar distribution on $SO(3)$, and $\mathbf{t}$ uniformly sampled from a $(-20 \dots 20) \times (-20 \dots 20) \times (-20 \dots 20)$ meters box, we produce a posed GHUM sample mesh $\mathbf{V}(\theta, \beta, \mathbf{r}, \mathbf{t})$. The associated markers can be retrieved by a simple (fixed) linear regression matrix $\mathbf{W} \in \mathbb{R}^{N_v \times N_m}$, such that $\mathbf{M} = \mathbf{W}\mathbf{V}(\theta, \beta, \mathbf{r}, \mathbf{t})$. In our experiments, we noticed that injecting noise at this point, *i.e.* $\mathbf{M} + \mathcal{N}(\mathbf{0}; \epsilon I)$, supports the more accurate retrieval of the full mesh given real, imprecise markers that one could find in motion capture datasets such as CMU or Human3.6M, or as produced by an image-based marker regressor. An overview of the poser function, denoted MP is given in fig. 2. We denote $\mathbf{V}_d$ the mesh that is directly predicted from markers. We denote $\mathbf{V}_p$ the mesh that is parametrically obtained by posing the GHUM model given parameters $(\widetilde{\theta}, \widetilde{\beta}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{t}})$ regressed from markers. For training, we use the loss

$$\mathcal{L} = \mathcal{L}_p(\mathbf{V}, \mathbf{V}_p) + \mathcal{L}_d(\mathbf{V}, \mathbf{V}_d), \tag{2}$$

where $\mathcal{L}_P$ and $\mathcal{L}_V$ are the mean-per-vertex errors, computed using a L_2 metric, between the input mesh, and the parametric and direct meshes, respectively. We also experimented with supervising $(\widetilde{\theta}, \widetilde{\beta}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{t}})$ directly, but learning was not successful. To make the training process easier, we subtract the mean marker position (computed as the 3d centroid of each $\mathbf{M}$) before regressing θ, β and $\mathbf{r}$ and we obtained lower reconstruction errors using this modification.

2.3. THUNDR

In fig. 3 we show an overview of our proposed hybrid learning architecture for monocular 3d body pose and shape estimation. Our architecture is different from existing pose and shape estimation methods, that directly regress the parameters of a human model (*i.e.* SMPL or GHUM) from a single feature representation of an image. We instead regress an intermediate 3d representation in the form of surface landmarks (markers) and regularize it in training using a statistical body model. Moreover, we preserve the spatial structure of high-level image features by avoiding pooling operations, and relying instead on self-attention to enrich our representation [33]. We draw inspiration from vision transformers [8], as we also use a hybrid convolutional-transformer architecture, and from [38], as we explore the idea of iteratively refining estimates by relying on cascaded, input-sensitive processing blocks, with homogeneous parameters.

Our network receives as input a cropped image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$ of a person, together with the pseudo ground-truth camera intrinsics $\mathbf{C}_c \in \mathbb{R}^{1 \times 4}$ of the crop (see § 2.1). We apply a convolutional neural network (CNN) on the input image and extract a downsampled feature map representation $\mathbf{F} \in \mathbb{R}^{\frac{W}{32} \times \frac{H}{32} \times D}$. We flatten the feature map along the spatial dimensions to get a sequence of $N = \frac{W}{32} \times \frac{H}{32}$ tokens. We append to each token the camera intrinsics and get our input feature sequence $\mathbf{F}_s \in \mathbb{R}^{N \times (D+4)}$. This sequence is linearly embedded by means of matrix $\mathbf{E} \in \mathbb{R}^{(D+4) \times D'}$, where D' is the embedding dimensionality, and concatenate it with an extra learnable [*markers*] token, $\mathbf{F}_{\text{m}} \in \mathbb{R}^{1 \times D'}$. Next, learnable positional embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D'}$ are added to the sequence. Different from standard transformer architectures, we use a single transformer encoder layer [33], *TL*, to iteratively refine our input representation for a number of L steps. We collect at each stage $l \in \{1 \dots L\}$, a

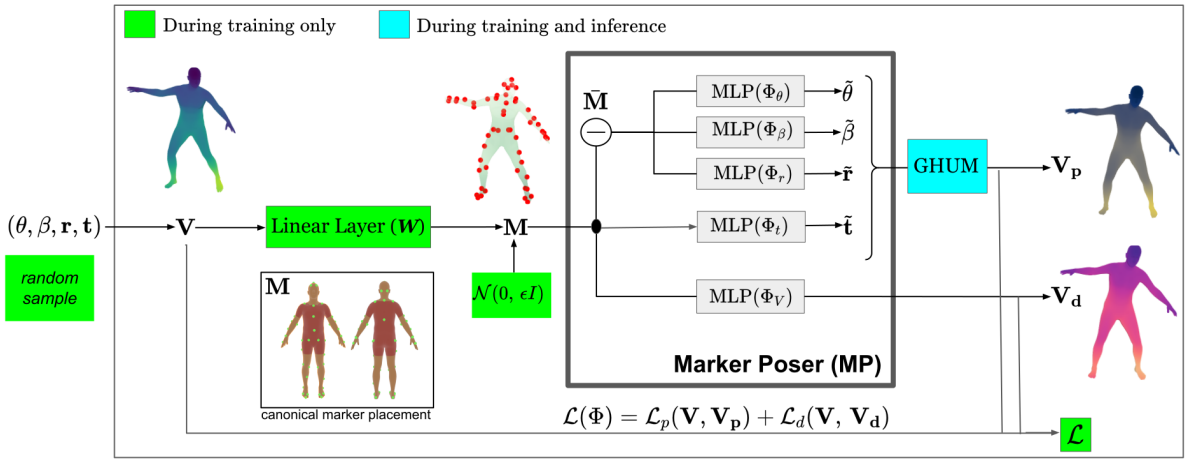

Figure 2: Our marker poser is based on a constrained marker-prediction pipeline which auto-encodes an initially generated, body mesh that is consistent with the human anthropometry $\mathbf{V}$ into a set of markers $\mathbf{M}$ via a linear layer characterized by a matrix $\mathbf{W}$. The markers are then used to predict both the GHUM parameters, resulting in a mesh $\mathbf{V}_p$ (we center the markers before regressing θ, β and $\mathbf{r}$) and a free-form mesh $\mathbf{V}_d$. Training losses ensure the consistency between $\mathbf{V}$, $\mathbf{V}_p$ and $\mathbf{V}_d$. We also show a detail of the cannonical marker placement, as attached on the GHUM model. Notice slight left/right and more pronounced front/back placement asymmetries that help disambiguate the model side and facing direction.

refinement update $\Delta\mathbf{M}_l \in \mathbb{R}^{N_m \times 3}$, with N_m the number of markers, from each transformed representation $\mathbf{Z}_l$ using a shared MLP applied on the representation of the [*markers*] token,

$$\mathbf{Z}_0 = \begin{bmatrix} \mathbf{F}_\mathrm{m} \\ \mathbf{F}_s\mathbf{E} \end{bmatrix} + \mathbf{E}_{pos} \tag{3}$$

$$\mathbf{Z}_l = TL(\mathbf{Z}_{l-1}) \tag{4}$$

$$\Delta\mathbf{M}_l = MLP(\mathbf{Z}_l^0). \tag{5}$$

The refinement updates $\Delta\mathbf{M}_l$ are added to the default marker coordinates, $\mathbf{M}_0$, as

$$\mathbf{M}_L = \mathbf{M}_0 + \lambda\Sigma_{l=1}^{L}\Delta\mathbf{M}_l, \tag{6}$$

where λ is a parameter controlling the step size. $\mathbf{M}_0$ are computed based on the default GHUM parameters, $(\theta_0, \beta_0, \mathbf{r}_0, \mathbf{t}_0)$, and camera intrinsics. That is, we find the optimal translation $\mathbf{t}_0^*$ such that the corresponding posed mesh projects in the center of the image [38]. Finally, $\mathbf{M}_0$ is computed as

$$\mathbf{M}_0 = \mathbf{W}\mathbf{V}(\theta_0, \beta_0, \mathbf{r}_0, \mathbf{t}_0^*). \tag{7}$$

We apply the pre-trained marker-based poser MP (see § 2.2) on $\mathbf{M}_L$ in order to recover the GHUM mesh and parameters, $\{\mathbf{V}_d, \widetilde{\theta}, \widetilde{\beta}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{t}}\}$. We also compute the mesh geometry using the standard GHUM poser from the regressed model parameters, $\mathbf{V}_p(\widetilde{\theta}, \widetilde{\beta}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{t}})$. During training, we use a mixed regime based on both weak 2d supervision losses and full 3d supervision losses, where data is available.

We include regularization losses for pose and shape, as

$$\mathcal{L}_{ps} = \|\widetilde{\beta}\|_2^2 + \|\widetilde{\theta}\|_2^2. \tag{8}$$

For this constraint to also affect the predicted markers $\mathbf{M}_L$ in a direct manner, we must formulate a consistency loss between the two representations. We set a novel loss that measures the *mean per-marker position error* (i.e. MPMPE) between the predicted markers and the markers on the surface of $\mathbf{V}_p$, *i.e.* $\mathbf{M}_p = \mathbf{W}\mathbf{V}_p$, as

$$\mathcal{L}_m = \frac{1}{N_m}\sum_{i=1}^{N_m}\|\mathbf{M}_L^i - \mathbf{M}_p^i\|_2. \tag{9}$$

We use a standard 2d reprojection loss measured with respect to either annotated or predicted keypoints, $\mathbf{j} \in \mathbb{R}^{K\times 2}$, weighted by a per-keypoint confidence score $\mathbf{s} \in \mathbb{R}^{K\times 1}$, with K the number of keypoints. From our directly regressed mesh $\mathbf{V}_d$ we extract 3d joints $\mathbf{J}$ via the standard GHUM regressor and project them using camera intrinsics $\mathbf{C}_c$ to predict 2d keypoints

$$\mathcal{L}_k = \frac{1}{K}\sum_{i=1}^{K}\mathbf{s}_i\|\mathbf{j}_i - \Pi(\mathbf{J}_i(\mathbf{V}_d), \mathbf{C}_c)\|_2. \tag{10}$$

Similarly to [38], we use a soft differentiable rasterizer [22] to compute a body part alignment loss with respect to either ground-truth or predicted body part maps $\mathbf{B} \in \mathbb{R}^{W\times H\times 15}$, with 15 different body part labels

$$\mathcal{L}_b = \frac{1}{W*H}\sum_{i=1}^{W*H}\|\mathbf{B}_i - R(\mathbf{V}_d, \mathbf{C}_c)_i\|_1, \tag{11}$$

where R is the rasterized image of the 3d body parts of $\mathbf{V}_d$, projected using camera intrinsics $\mathbf{C}_c$.

Given access to 3d supervision with ground-truth vertices $\mathbf{V}_{gt}$ and joints $\mathbf{J}_{gt}$, we use standard vertex and 3d keypoints losses:

$$\mathcal{L}_f = \lambda_v \mathcal{L}_v(\mathbf{V}_d, \mathbf{V}_{gt}) + \lambda_j \mathcal{L}_j(\mathbf{J}, \mathbf{J}_{gt}),$$

with $\mathcal{L}_v$ the MPVE (mean per vertex error) metric and $\mathcal{L}_j$ the MPJPE (mean per joint position error) metric. Parameters λ_v and λ_j control the importance of each loss.

Finally, we can write our full loss function, as follows

$$\mathcal{L} = \lambda_{ps}\mathcal{L}_{ps} + \lambda_m \mathcal{L}_m + \lambda_k \mathcal{L}_k + \lambda_b \mathcal{L}_b + \mathcal{L}_f \qquad (12)$$

where λ are used to weigh the different loss components. The fully supervised loss $\mathcal{L}_f$ is only used if there exists 3d ground truth information.

3. Experiments

Datasets We use two datasets containing images in-the-wild, COCO2017 [21] (30,000 images) and OpenImages [18] (24,000 images) for our weakly-supervised training (WS). We use the 2d keypoint annotations where available, otherwise we rely on a 2d pose detector to supplement missing annotations and use an additional confidence score per keypoint.

For the fully-supervised (FS) experiments, we use two standard datasets Human3.6M [12] and 3DPW [34]. Because the ground-truth of 3DPW is provided as SMPL [23] 3d meshes, we use GHUM fits to these meshes to report the vertex-to-vertex errors. The MPJPE metrics are reported on the 3d joints regressed from the ground-truth SMPL meshes, as standard in the literature. Differently from existing methods, we use less 3d supervision, with superior results. We did not include additional datasets such as MuCo-3DHP [25], MPI-INF-3DHP [26] or UP3D [19], but we believe they could be helpful in further increasing our reconstruction performance.

Implementation details In all our experiments we use a ResNet50 [11] backbone pretrained for the ImageNet [7] image classification task. Our complete architecture has 25M parameters, 23.5M for the backbone and 1.5M for the transformer layer and the MLP regressor. We use $L = 4$ stages, step size $\lambda = 0.1$, an embedding size 256 and 8 heads for the MultiHeadAttention layer. We train the network for 50 epochs, with batch size of 32, base learning rate of $1e-4$ and exponential decay 0.99. Our marker poser MP has 8.5M parameters and consists of MLPs with a hidden layer size of 256. The network is trained for 1M steps with a batch size of 128. All our networks were trained on a single V100 GPU with 16GB of memory. Our code is implemented in TensorFlow.

3.1. 3D Pose and Shape Reconstruction

For this task, we report several common error metrics that are used for evaluating the error of 3d reconstruction. Most commonly used for 3d joint errors are mean per joint position error (MPJPE) and MJPE-PA, which is MPJPE after rigid alignment of the prediction with ground truth via Procrustes Analysis. The latter metric, removes global misalignment (*i.e.* scale and rotation) and mainly evaluates the quality of the reconstructed 3d pose. For evaluating 3d shape we use the MPVPE metric between the vertices of the predicted and ground-truth meshes, respectively.

We evaluate our networks on the two datasets that provide 3d ground-truth information, Human3.6M and 3DPW. For the Human3.6M dataset, there are three commonly used evaluation protocols in the literature. Protocols P1 and P2 consider splitting the official training set into new training and testing subsets, with subjects S1, S5-S8 for training and S9 and S11 for evaluation. P1 evaluates on all available camera views in testing, while P2 only on a single predefined camera view (we consider this to be a highly inconclusive protocol due to its small size and design but report on it in order to compare to other methods). The third and most representative protocol we consider is the official one, where we evaluate on the hold-out test dataset of 900K samples. We also submit predictions on the official website for other methods (where code and models are available) to get comparable results. To be fair in our comparison with other methods, we do not retrain on the whole official training dataset. We show results for all protocols in tables 1, 3 and 5. For P1, we report results for both the weakly supervised regime (WS) and for the mixed regime (WS+FS) in order to compare with prior work. For the official protocol, we report only the MPJPE rounded to the nearest integer, as this is the format the results are returned by the official site. On all protocols and in all training regimes, we obtain state-of-the-art results. In table 4 we report errors on the testing split of the 3DPW dataset. We obtain state-of-the-art results, in both the WS+FS regime and the WS regime.

In fig. 4 we show qualitative reconstructions from THUNDR in-the-wild where one can observe that direct mesh reconstructions $\mathbf{V}_d$ have better image alignment in general. We also show the image attention map for the [*markers*] token, aggregated over all transformer layers. Notice how the network learns to focus on faces, hands and feet.

Ablation Studies In table 2, we ablate different methodological choices in our proposed architecture in the weakly supervised regime and report results on protocol P2 of Human3.6M dataset. First, we change THUNDR to directly regress GHUM parameters (*i.e.* $\widetilde{\beta}, \widetilde{\theta}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{t}}$) from the input image, skipping our intermediate marker representation and removing the marker poser MP. The convolutional-transformer architecture stays mostly the same, with some minor mod-

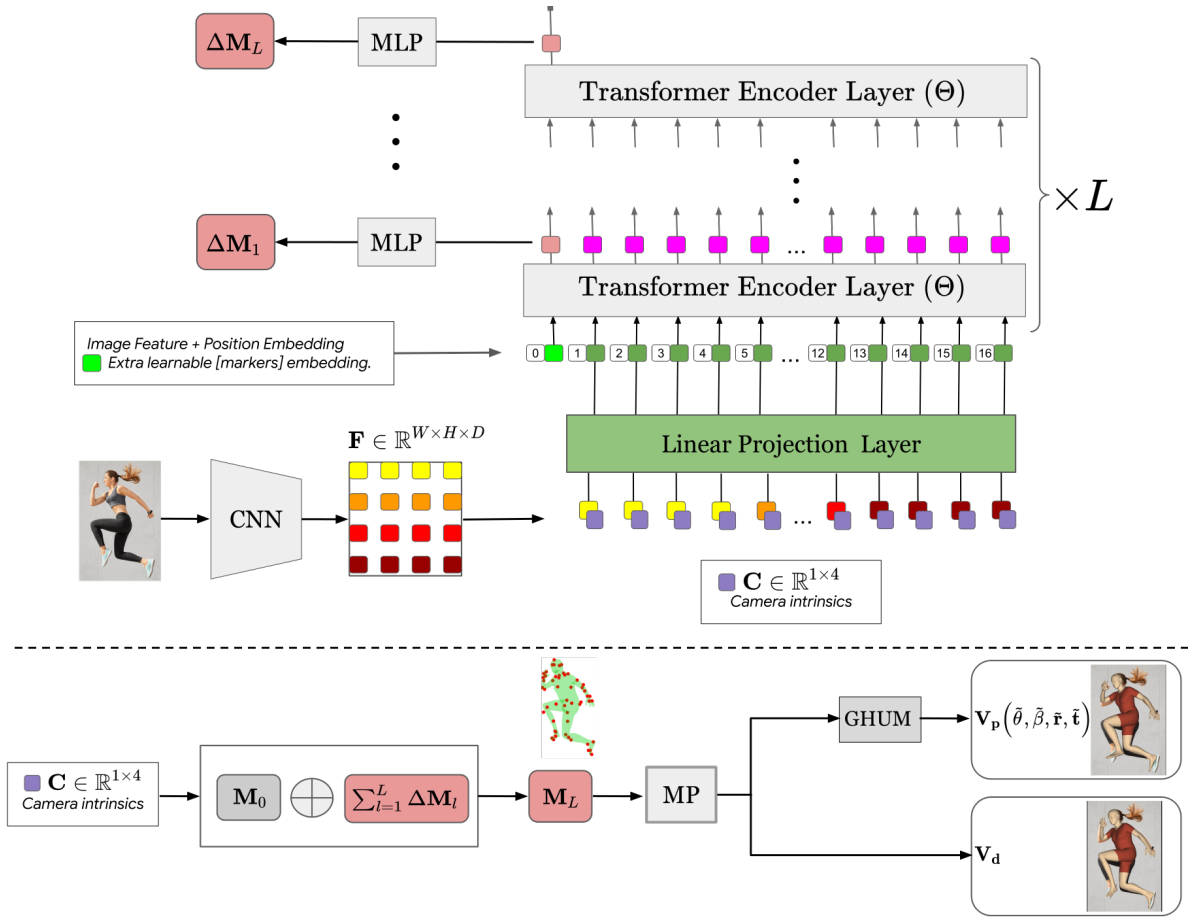

Figure 3: Overview of our proposed THUNDR architecture, to estimate the parameters of a generative human model (GHUM). *(Top)* Given an input image, we first use a CNN to extract a feature map $\mathbf{F} \in \mathbb{R}^{W\times H\times D}$, where W and H represent the spatial extent, and D the number of channels per feature. In this example $W = H = 4$. We serialize the feature map and concatenate to each feature the camera intrinsics of the image, $\mathbf{C}$. Next, we take our sequence, linearly embed it and add positional encoding. We also add an extra learnable [*markers*] token to the input. This representation is iteratively transformed L times through *the same* transformer encoder layer with learnable weights Θ. At each transformation stage l, we gather the representation of the [*markers*] token, feed it through an MLP and regress the marker coordinates refinement $\Delta\mathbf{M}_l$. *(Bottom)* We compute the default marker coordinates $\mathbf{M}_0$ as a function of the image camera intrinsics and default GHUM model parameters. The regressed marker coordinates displacements are added to it and the result represents the final estimated marker coordinates $\mathbf{M}_L$. We use the pre-trained marker-based poser MP to get our predicted GHUM model vertices and parameters.

Method	MPJPE-PA	MPJPE	Translation Error
HMR (WS) [15]	67.45	106.84	NR
HUND (SS) [38]	66.0	102	175.0
THUNDR (WS)	**62.2**	**87.0**	161.6
HMR [15]	58.1	88.0	NR
HUND [38]	53.0	72.0	160.0
THUNDR	**39.8**	**55.0**	**143.9**

Table 1: Performance of different pose and shape estimation methods on the Human3.6M dataset, with training/testing under protocol P1.

ifications to accommodate more output variables (*i.e.* we use 4 extra input tokens, one for each GHUM parameter, instead of 1). This performs worse than our proposed architecture THUNDR and this shows that our intermediate marker representation is easier to learn from image features. Next, as in our full method we only use the direct mesh $\mathbf{V}_d$, we also show the errors if we instead evaluate on the parameteric mesh $\mathbf{V}_p$ (we denote this by THUNDR-$\mathbf{V}_p$).

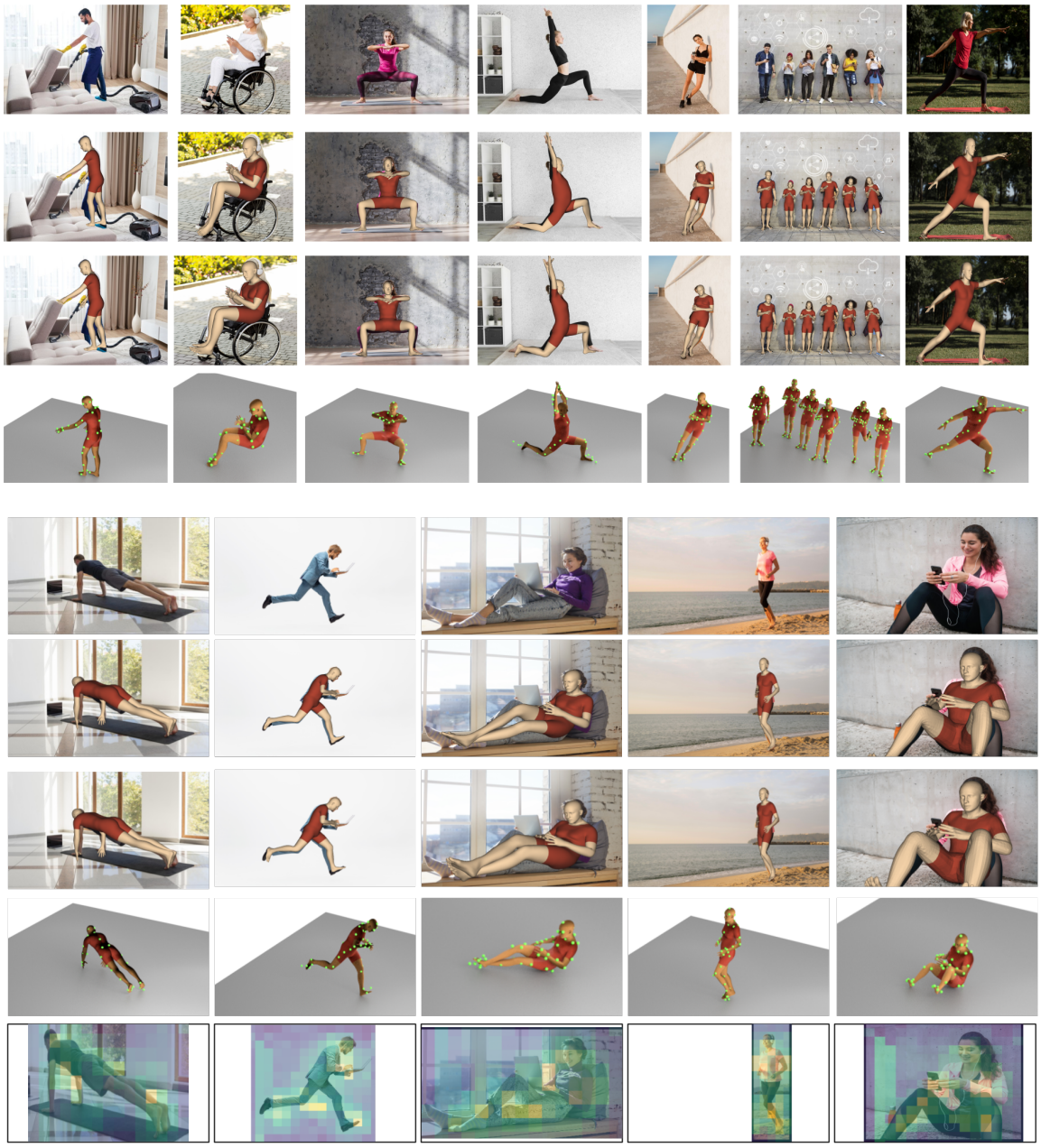

Figure 4: Results of THUNDR on images in the wild. From top to bottom: *(i)* input image *(ii)* direct mesh reconstructions $\mathbf{V}_d$ *(iii)* parameteric mesh reconstructions $\mathbf{V}_p$. Notice that direct mesh reconstruction aligns better, particularly the feet and the limbs. *(iv)* reconstructions seen from a different viewpoint with regressed marker representations shown in green. *(v)* the image attention map for the [*markers*] token, aggregated over all transformer layers.

These results are also better than THUNDR-GHUM, but worse than THUNDR. This again suggests the utility of our intermediate representation and the importance of working with two separate mesh reconstructions.

Marker Poser We present more details on the training of the marker poser and its additional benefits, outside of the transformer-based 3d pose reconstruction architecture. During training, we experiment with 4 levels of Gaussian

Method	MPJPE-PA	MPJPE
THUNDR-GHUM (WS)	63.5	95.4
THUNDR-$\mathbf{V}_p$ (WS)	61.8	88.3
THUNDR (WS)	**59.7**	**83.4**

Table 2: Ablation study on different variations of THUNDR: THUNDR-GHUM directly regresses GHUM parameters from the image and THUNDR-$\mathbf{V}_p$ is our standard version were we instead evaluate on the predicted parameteric mesh $\mathbf{V}_p$. This evaluation is done in a weakly supervised regime and we report error metrics on Human3.6M protocol P2.

noise added to the markers, as $\epsilon \in \{0, 20, 50, 100\}$ mm. We ablate each one of the trained marker poser models on the Human3.6M ground-truth marker data. The best performance in reconstruction is achieved for the network with $\epsilon = 50$ mm. This model is used in all of our other experiments. During training, the error on the direct mesh reconstruction reaches 25 mm MPVPE, while the parameteric mesh reconstruction reaches 37 mm MPVPE.

Mesh Fitting We test our marker-based poser on the Human3.6M dataset, for which the authors shared 3D marker positions for the training data. We fit an associated GHUM mesh in two ways: *(i)* by minimizing an energy that takes into account 3d marker ground truth, 2d reprojection errors for all GHUM 3d body joints (including hands and face) and a semantic alignment cost, and *(ii)* by simply running our trained marker poser on the ground-truth 3d marker positions to produce a mesh $\mathbf{V}_d$. For a sequence fitting example, see our Sup. Mat. First, we compute the mean per-marker error for the models $\mathbf{V}_{gt}$ obtained from energy optimization to ground-truth makers $\mathbf{M}_{gt}$ (*i.e.* those recovered from mocap data). This gives an error of 38.4 mm, with an average processing rate of 0.15 frames/second. Second, we compute the errors of markers placed on the predicted mesh $\mathbf{V}_d$ given ground-truth marker positions. This achieves a slightly higher error of 44.3 mm, but with an average processing rate of 1000 frames/second, when ran sequentially. Note that our marker poser has never seen the marker sequences of Human3.6M during training, as the marker poser was trained with samples drawn from a normalizing flow prior based on the CMU motion capture dataset [1].

Method	MPJPE-PA	MPJPE
HMR [15]	56.8	88.0
GraphCMR [17]	50.1	NR
Pose2Mesh [5]	47.0	64.9
I2L-MeshNet [27]	41.1	55.7
SPIN [16]	41.1	NR
METRO [20]	36.7	54.0
THUNDR	**34.9**	**48.0**

Table 3: Performance of different pose and shape estimation methods on the Human3.6M dataset, protocol P2.

Method	MPJPE-PA	MPJPE	MPVPE
HUND [38] (SS)	70.3	98.1	NR
THUNDR (WS)	**59.9**	**86.8**	NR
HMR [15]	81.3	NR	NR
GraphCMR [17]	70.2	NR	NR
SPIN [16]	59.2	NR	116.4
Pose2Mesh [5]	58.9	89.2	NR
I2L-MeshNet [27]	57.7	93.2	NR
HUND [38]	56.5	87.7	NR
METRO [20]	**47.9**	77.1	88.2
THUNDR	51.5	**74.8**	***88.0**

Table 4: Performance of different pose and shape estimation methods on the 3DPW dataset.**Shape evaluation is done on GHUM.*

Method	MPJPE
HMR [15]	89
SPIN [16]	68
HUND [38]	66
THUNDR	**53**

Table 5: Performance of different methods on the Human3.6M official, representative held-out test set, containing 900K samples.

Ethical Considerations Our methodology aims to decrease bias by introducing flexible forms of self-supervision which would allow, in principle, for system bootstrapping and adaptation to new domains and fair, diverse subject distributions, for which labeled data may be difficult or impossible to collect upfront. Applications like visual surveillance and person identification would not be effectively supported currently, given that model's output does not provide sufficient detail for these purposes. This is equally true of the creation of potentially adversely-impacting deepfakes, as we do not include an appearance model or a joint audio-visual model.

4. Conclusions

We have presented THUNDR, a transformer-based deep neural network methodology to reconstruct the 3d pose and shape of people, given monocular RGB images. Faced with the difficult issue of handling not directly observable human body joints, on which nevertheless many error metrics are based, and aiming at both reconstruction accuracy and good self-supervised learning and generalization under anthropometric human body constraints, we propose a novel model that combines a surface-marker representation with 3d statistical body regularization. The model is designed around a learnable pipeline that refines multiple transformer layers for computational efficiency and for precise, task-sensitive, image feature localization. We demonstrate state-of-the-art results on Human3.6M and 3DPW, in both the fully-supervised and the self-supervised regimes.

References

[1] Carnegie Mellon Motion Capture Database. http://mocap.cs.cmu.edu. . 8

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, pages 3395–3404, 2019. 2

[3] Benjamin Biggs, Sébastien Ehrhadt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *arXiv preprint arXiv:2011.00980*, 2020. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787. Springer, 2020. 2, 8

[6] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, pages 768–784. Springer, 2020. 2

[10] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. 5

[12] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014. 5

[13] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, pages 5243–5252, 2020. 2

[14] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 2

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 6, 8

[16] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 2, 8

[17] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 8

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 5

[19] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[20] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 2, 8

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[22] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, pages 7708–7717, 2019. 4

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 2015. 2, 5

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 2

[25] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, pages 120–130. IEEE, 2018. 5

[26] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 2017. 5

[27] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 8

[28] A.I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *CVPR*, 2017. 2

[29] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, September 2018. 2

[30] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *IJRR*, 22(6):371–393, 2003. 2

[31] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 2

[32] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[34] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 5

[35] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. *CVPR*, 2020. 2

[36] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019. 2

[37] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *ECCV*, 2020. 2

[38] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. *CVPR*, 2021. 2, 3, 4, 6, 8

[39] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 2

[40] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *CVPR*, pages 7054–7063, 2020. 2

[41] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, pages 7376–7385, 2020. 2

[42] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, pages 3372–3382, 2021. 2

[43] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *arXiv preprint arXiv:1812.07035*, 2018. 2